ANALYSIS

30(135), April 1, 2015

ISSN 2278–5469 EISSN 2278–5450

Discovery

# A survey on data mining trends, applications and techniques

**Saraswathi K[1], Ganesh Babu V[2]**

1.Assistant Professor (Research Scholar), Department of CS, Nehru Memorial College, Trichy, India; Email: saras_shan@yahoo.com
2.Assistant Professor, Department of CS, Government College for Women, Mandya, India; Email: vgbzone@gmail.com

**Publication History**
Received: 18 February 2015
Accepted: 22 March 2015
Published: 1 April 2015

**Citation**
Saraswathi K, Ganesh Babu V. A survey on data mining trends, applications and techniques. *Discovery*, 2015, 30(135), 383-389



**General Note**
Article is recommended to print as color digital version in recycled paper.

## ABSTRACT

In this paper, the overview of the data mining applications and its trends are surveyed. This paper imparts more number of applications of the data mining. The features of Data Mining Systems are specified here. This paper also summarizes the list of techniques used in data mining.

*Index Terms*: Data Mining, Data Mining Trends, Data Mining Applications, Data Mining Techniques, Data Mining Features.

## 1. INTRODUCTION

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. “Data mining sometimes called data or knowledge discovery. Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it,

and summarize the relationships identified. This paper summarizes some list of trends in data mining such as Application exploration, Scalable data mining methods, Integration of data mining, Data mining and software engineering, Visual Data Mining, Biological data mining, Web mining, Hypertext and Hyper Media Data Mining, Distributed Data mining, Real-time data mining, and Multi-database data mining. This paper also summarizes the Applications and Techniques of data mining.

## 2. DATA MINING TRENDS

As different types of data are available for data mining tasks, data mining approaches poses many challenging research issues in data mining. The design of a standard data mining languages, the development of effective and efficient data mining methods and systems, the construction of interactive and integrated data mining environments, and the applications of data mining to solve large applications problems are important tasks for data mining researches and data mining system and application developers. This paper reviews some of the trends in data mining.

## 3. DATA MINING APPLICATIONS

### *A. Financial Data Analysis*

The financial data in banking and financial industry are generally reliable and of high quality which eases the systematic data analysis and data mining.

### *B. Retail Industry*

The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That leads to improve the quality of customer service and good customer preservation and satisfaction.

### *C. Telecommunication Industry*

Data Mining in Telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service.

### *D. Biological Data Analysis*

Biological data mining is a very important part of Bioinformatics. Now a days we see that there is vast growth in the field of Biology such as genomics, proteomics, functional Genomics and biomedical research.

### *E. Intrusion Detection*

Intrusion refers to any kind of action that threatens integrity, confidentiality, or availability of network resources. Intrusion detection is a critical component of network administration.

### *F. Data Mining Applications in Sales/Marketing*

Data mining is used for market basket analysis to provide information on what product combinations were purchased together, when they were bought and in what sequence. This information helps businesses to promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products that they may have been missed or overlooked. Retail Company's uses data mining to identify customer's behavior buying patterns.

### *G. Data Mining Applications in Banking and Finance*

Credit card expenses by customer groups can be identified by using data mining. The hidden correlations between different financial indicators can be discovered by using data mining. From historical market data, data mining enables to identify stock trading rules.

### *H. Data Mining Applications in Insurance*

Data mining is applied in claims analysis such as identifying which medical procedures are claimed together. Data mining enables to forecasts which customers will potentially purchase new policies. Data mining allows insurance companies to detect risky customers' behavior patterns. Data mining helps to detect fraudulent behavior.

### *I. Data Mining Applications in Transportation*

Data mining helps to determine the distribution Schedules among warehouses and outlets and analyze loading patterns.

**J. Data Mining Applications in Medicine**

Data mining enables to characterize patient activities to see incoming office visits. Data mining helps to identify the patterns of successful medical therapies for different illnesses.

***K. Other Scientific Applications***

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy etc. There is large amount of data sets being generated because of the fast numerical simulations in various fields such as climate, and ecosystem modeling, chemical engineering, fluid dynamics etc. The applications of data mining in field of Scientific Applications are Data Warehouses and data preprocessing, Graph-based mining, Visualization and domain specific knowledge.

## 4. SCALABLE DATA MINING METHODS

Classification is an important problem in the emerging field of data mining. Although classification has been studied extensively in the past, most of the classification algorithms are designed only for memory-resident data, thus limiting their suitability for data mining large data sets. SLIQ is a decision tree classifier that can handle both numeric and categorical attributes. It uses a novel pre-sorting technique in the tree-growth phase. This sorting procedure is integrated with a breadth-first tree growing strategy to enable classification of disk-resident datasets. SLIQ also uses a new tree-pruning algorithm that is inexpensive, and results in compact and accurate trees. The combination of these techniques enables SLIQ to scale for large data sets and classify data sets irrespective of the number of classes, attributes, and examples (records), thus making it an attractive tool for data mining.

## 5. INTEGRATION OF DATA MINING

Data mining techniques, based on statistics and machine learning can significantly boost the ability to analyze data. Despite the potential effectiveness of data mining to significantly enhance data analysis, this technology is destined be a niche technology unless an effort is made to integrate this technology with traditional database systems. This is because data analysis needs to be consolidated at the warehouse for data integrity and management concerns. Therefore, one of the key challenges is to enable integration of data mining technology seamlessly within the framework of traditional database systems.

***Data Mining in SQL Server 2000***

Microsoft SQL Server 2000 integrates for the first time Data Mining capabilities together with relational and OLAP database engines. The Analysis Services component of SQL Server 2000 contains a data-mining engine that is exposed through an OLE DB for DM interface. The data-mining engine is integrated in both the server component of analysis services, which includes also the OLAP engine, as well as on the client component. The data-mining engine provides two classes of scalable algorithms based on work done in Microsoft Research such as classification and segmentation (clustering). Future releases of the product will extend the repertoire of algorithms. The product provides a full GUI based console for the administration of the data-mining model including the creation, training, browsing of content and access security management. While the detailed aspects of the UI are not the central focus of this paper, the UI serves as a natural way to introduce a user to how data mining can be done naturally within the SQL Server framework. However, it should be noted that the primary contribution of our proposal is the backend components that support this UI and understand the semantics of data mining. Creation of data mining model is done through the "Mining Model Wizard" that guides the user through four easy steps for building the basic data mining model: the selection of the table containing the cases, the selection of the algorithm (classification/clustering), the identification of the case-key column and the selection of the predictable columns and the input columns.

## 6. DATA MINING AND SOFTWARE ENGINEERING

The first step in the knowledge discovery process is to gain understanding about the data that is available and the business goals that drive the process. This is essential for software engineering data mining endeavors, because unavailability of data for mining is a factor that limits the questions which can be effectively answered. This section, describes software engineering data that are available for data mining and analysis. Current software development processes involve several types of resources from which software-related artifacts can be obtained. Software 'artifacts' are a product of software development processes. Artifacts are generally loss and thus cannot provide a full history or context, but they can help piece together understanding and provide further insight. There are many data sources in software engineering. In this paper, we focus only on four major groups and describe how they may be used for mining software engineering data.

First, the vast majority of collaborative software development organizations utilize revision control software1 (e.g., CVS, Subversion, etc.) to manage the ongoing development of digital assets that may be worked on by a team of people. Such systems maintain a historical record of each revision and allow users to access and revert to previous versions. By extension, this provides a way to analyze historical artifacts produced during software development, such as number of lines written, authors which wrote particular lines or any number of common software metrics.

Second, most large organizations (and many smaller ones) also use a system for tracking software defects. Bug tracking software (such as Bugzilla, JIRA, FogBugz, etc.) associates bugs with meta-information (status, assignee, comments, dates and milestones, etc.) that can be mined to discover patterns in software development processes, including the time-to-fix, defect-prone components, problematic authors, etc. Some bug trackers are able to correlate defects with source code in a revision system. 246 *Q. Taylor and C. Giraud-Carrier[8].*

Third, virtually all software development teams use some form of electronic communication (e-mail, instant messaging, etc.) as part of collaborative development (communication in small teams may be primarily or exclusively verbal, but such cases are inconsequential from a data mining perspective). Text mining techniques can be applied to archives of such communication to gain insight into development processes, bugs and design decisions.

Fourth, software documentation and knowledge bases can be mined to provide further insight into software development processes. This approach is useful to organizations that use the same processes across multiple projects and want to examine a process in terms of overall Effectiveness or fitness for a given project.

## 7. VISUAL DATA MINING

Visual data mining is an effective way to discover knowledge from huge amounts of data. A visual data mining system must be syntactically simple to be useful. Simple to learn means use of intuitive and friendly input mechanisms as well as instinctive and easy-to-interpret output knowledge. Simple to apply means an effective discourse between humans and information. Simple to retrieve means a customized data structure to facilitate fast and reliable searches. Simple to execute means a minimum number of steps needed to achieve the results. A reliable visual data mining system must provide estimated error or accuracy of the projected information for each step of the mining process. A reusable visual data mining system must be adaptable to a variety of systems and environments to reduce the customization effort, provide assured performance, and improve system portability. A practical visual data mining system must be generally and widely available. The quest for new knowledge or deeper insights of existing knowledge cannot be planned. It may mean a portable system through telephone links or an embedded (local) system within the information domain. Finally, a complete visual data mining system must include security measures to protect the data, the newly discovered knowledge, and the user's identity because of various social issues.

## 8. DISTRIBUTED DATA MINING

One area of data mining which is attracting a good amount of attention is that of distributed and collective data mining. Much of the data mining which is being done currently focuses on a database or data warehouse of information which is physically located in one place. However, the situation arises where information may be located in different places, in different physical locations. This is known generally as distributed data mining (DDM). Therefore, the goal is to effectively mine distributed data which is located in heterogeneous sites. Examples of this include biological information located in different databases, data which comes from the databases of two different firms, or analysis of data from different branches of a corporation, the combining of which would be an expensive and time-consuming process. Distributed data mining (DDM) is used to offer a different approach to traditional approaches analysis, by using a combination of localized data analysis, together with a "global data model." In more specific terms, this is specified as: -performing local data analysis for generating partial data models, and -combining the local data models from different data sites in order to develop the global model. This global model combines the results of the separate analyses. Often the global model produced, especially if the data in different locations has different features or characteristics, may become incorrect or ambiguous. This problem is especially critical when the data in distributed sites is heterogeneous rather than homogeneous. These heterogeneous data sets are known as vertically partitioned datasets.

## 9. REAL TIME DATA MINING

Many applications involving stream data (such as e-commerce, Web mining, stock analysis, intrusion detection, mobile data mining, and data mining for counterterrorism) require dynamic data mining models to be built in real time. Additional development is needed in this area. E-commerce is also the most prospective domain for data mining. It is ideal because many of the ingredients required for successful data mining are easily available: data records are plentiful, electronic collection provides reliable data, insight

can easily be turned into action, and return on investment can be measured. The integration of commerce and data mining significantly improve the results and guide the users in generating knowledge and making correct business decisions. This integration effectively solves several major problems associated with horizontal data mining tools including the enormous effort required in pre-processing of the data before it can be used for mining, and making the results of mining actionable.

## 10. BIOLOGICAL DATA MINING

Biological data mining is the activity of finding significant information in bimolecular data. The significant information may refer to motifs, clusters, genes and protein signature. Although biological data mining can be considered under "application exploration" or "mining complex types of data," the unique combination of complexity, richness, size and importance of biological data warrants are special attention in data mining.

## 11. WEB MINING

The development of World Wide Web and its usage grows, it will continue to generate ever more content, structure, and usage data and the value of Web mining will keep increasing. Research needs to be done in developing the right set of Web metrics, and their measurement procedures, extracting process models from usage data, understanding how different parts of the process model impact various Web metrics of interest, how the process models change in response to various changes that are made-changing stimuli to the user, developing Web mining techniques to improve various other aspects of Web services, techniques to recognize known frauds and intrusion detection.

## 11. HYPERTEXT / HYPERMEDIA DATA MINING

The hypertext and hypermedia data is a collection of data from online catalogues, digital libraries, and online information data bases which include hyperlinks, text markups and other forms of data. Web mining is the application of data mining to discover the patterns from the Web. The important data mining technique used for hypertext and hypermedia data are Classification (supervised learning), Clustering (unsupervised learning).

## 12. MULTIMEDIA DATA MINING

The multimedia data includes images, video, audio, and animation. The data mining techniques that are applied on multimedia data are rule based decision tree classification algorithms like Artificial Neural Networks, Instance-based learning algorithms, Support Vector Machines, also association rule mining, clustering methods.

## 13. MULTI-DATABASE DATA MINING

Data mining algorithms look for patterns in data. While most existing data mining approaches look for patterns in a single data table, multi-relational data mining (MRDM) approaches look for patterns that involve multiple tables (relations) from a relational database. In recent years, the most common types of patterns and approaches considered in data mining have been extended to the multi-relational case and MRDM now encompasses multi-relational (MR) association rule discovery, MR decision trees and MR distance-based methods, among others. MRDM approaches have been successfully applied to a number of problems in a variety of areas, most notably in the area of bioinformatics.

## 14. DATA MINING TECHNIQUES

### *1. Regression analysis*

Regression models are the mainstay of predictive analytics. The linear regression model analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. That relationship is expressed as an equation that predicts the response variable as a linear function of the parameters.

### *2. Choice modeling*

Choice modeling is an accurate and general-purpose tool for making probabilistic predictions about decision-making behavior. It behooves every organization to target its marketing efforts at customers who have the highest probabilities of purchase. Choice models are used to identify the most important factors driving customer choices. Typically, the choice model enables a firm to compute an individual's likelihood of purchase, or other behavioral response, based on variables that the firm has in its database, such as geo-demographics, past purchase behavior for similar products, attitudes, or psychographics.

***3. Rule induction***

Rule induction involves developing formal rules that are extracted from a set of observations. The rules extracted may represent a scientific model of the data or local patterns in the data. One major rule-induction paradigm are the association rule. Association rules are about discovering interesting relationships between variables in large databases. It is a technique applied in data mining and uses rules to discover regularities between products. For example, if someone buys peanut butter and jelly, he or she is likely to buy bread. The idea behind association rules is to understand when a customer does X, he or she will most likely do Y. Understanding those kinds of relationships can help with forecasting sales, promotional pricing, or product placements.

***4. Network/Link Analysis***

This is another technique for associating like records. Link analysis is a subset of network analysis. It explores relationships and associations among many objects of different types that are not apparent from isolated pieces of information. It is commonly used for fraud detection and by law enforcement.

***5. Clustering/Ensembles***

Cluster analysis, or clustering, is a way to categorize a collection of "objects," such as survey respondents, into groups or clusters to look for patterns. Ensemble analysis is a newer approach that leverages multiple cluster solutions (an ensemble of potential solutions). There are various ways to cluster or create ensembles. Regardless of the method, the purpose is generally the same—to use cluster analysis to partition into a group of segments and target markets to better understand and predict the behaviors and preferences of the segments. Clustering is a valuable predictive-analytics approach when it comes to product positioning, new-product development, usage habits, product requirements, and selecting test markets.

***6. Neural networks***

Neural networks were designed to mimic how the brain learns and analyzes information. Organizations develop and apply artificial neural networks to predictive analytics in order to create a single framework. The idea is that a neural network is much more efficient and accurate in circumstances where complex predictive analytics is required, because neural networks comprise a series of interconnected calculating nodes that are designed to map a set of inputs into one or more output signals. Neural networks are ideal for deriving meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by humans or other computer techniques. Marketing organizations find neural networks useful for predicting customer demand and customer segmentation.

***7. Memory-based reasoning (MBR)/Case-based reasoning.***

This technique has results similar to a neural network's but goes about it differently. MBR looks for “neighbor" kind of data rather than patterns. It solves new problems based on the solutions of similar past problems.MBR is an empirical classification method and operatesby comparing new unclassified records with known examples and patterns.

***8. Decision trees***

Decision trees use real data-mining algorithms to help with classification. A decision-tree process will generate the rules followed in a process.Decision trees are useful for helping you choose among several courses of action and enable you to explore the possible outcomes for various options in order to assess the risk and rewards for each potential course of action.

***9. Uplift modeling.***

This technique directly models the incremental impact of targeting marketing activities. The uplift of a marketing campaign is usually defined as the difference in response rates between a treated group and a randomized control group. Uplift modeling uses a randomized scientific control to measure the effectiveness of a marketing action.

**15. CONCLUSION**

Data mining is to discover or extract knowledge or data from large amount of database. This paper introduces briefly reviewed the concept of data mining, areas of data mining, data mining applications and its techniques where used today. It would be helpful to researchers to focus on the various trends of data mining. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

## FUTURE TRENDS

The complexity of data mining must be hidden from end-users before it will take the true center stage in an organization. Business use cases can be designed, with tight constrains, around data mining algorithms. Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments.

## REFERENCE


1. Gorunescu, F, Data *Mining: Concepts, Models, and Techniques*, Springer, 2011.
2. http://www.tutorialspoint.com
3. http://www.zentut.com
4. http://www.marketingprofs.com
5. Anvik, J. (2006) 'Automating bug report assignment', in *Proceedings of the 28th International Conference on Software Engineering*, pp.937–940.
6. J. HUYSMANS, B. BAESENS, D. MARTENS,K. DENYS and J.VANTHIENEN "New Trends in Data Mining" Tijdshrift voor Economie en Management Vol. L, 4, 2005
7. Vikas Gupta, Prof. Devanand "Tools, Techniques, Applications, Trends and Issues International Journal of Scientific & Engineering Research Volume 4, Issue 3, March-2013
8. Amit Kapoor. *International Journal of Emerging Trends & Technology in Computer Science,* Volume 3, Issue 1, January – February 2014 "Data Mining: Past, Present and Future Scenario"
9. Surajit Chaudhuri: Data Mining and Database Systems: Where is the Intersection? Data Engineering Bulletin 21(1)
10. Breimanet. al. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.